Eur. J. Biochem. 83, 1–3 (1978)

Chromosome Structure and Function. Future Prospects

Francis H. C. CRICK

Salk Institute, San Diego, California

It is clear from the two symposia and the workshop on "The Structure and Function of Chromatin" that there has been a big advance in our understanding of the three dimensional organization of chromosomes at the first level of coiling of the DNA, due mainly to the concept of the core nucleosome particle (or 'platysome'), although the exact location of histone H1 and the arrangement of the linker regions between platysomes are still in doubt. Even less is known about the precise location of the non-histone proteins and the details of the higher levels of coiling. Some 10 to 20% of the nucleosomes (the amount depends on the tissue under study) have a looser configuration which makes them more sensitive to the nuclease DNase I. These 'active' nucleosomes appear to include much of those stretches of DNA which are being transcribed in any particular tissue.

In addition several lines of work suggest that (as claimed for *Escherichia coli*) the DNA in eukaryotes is arranged in 'domains'. The average size of these domains is estimated, in very round terms, to be about 50000 base-pairs but the distribution of sizes about the average is as yet unknown. Nor is the exact nature and functional significance of these domains at all clear though there is no lack of informed guesses on this point. In particular one would like to know whether the nucleosomes in a single domain are, at any one time, all in the 'active' state or all in the more compact inactive state or whether, on the other hand, nucleosomes of both types occur at the same time in a single domain.

During this recent period there has also been a very big improvement in our ability to study the 1 D structure of DNA; that is, the nucleotide sequence. This has come from the well-known advances in genetic engineering which enable longish stretches of DNA (of the order of 10^4 base-pairs) to be inserted into microorganisms and thus multiplied up so that biochemically useful amounts of the DNA of pure eukaryotic 'genes' can be obtained for further study. The use of restriction enzymes and hybridization techniques has allowed the rapid mapping, on a relatively coarse scale, of these DNA segments, while two extremely fast methods are available for obtaining exact nucleotide sequences. We can expect that in the next few years this detailed information about DNA sequences will grow from its present modest flow into an almost unmanageable flood. Special arrangements, probably involving some central computing facility, will almost certainly have to be made to collect and store these sequences and to distribute them to all interested workers in the field.

What will such sequence information tell us? Already it has yielded unexpected results. Sequences coding for one particular mRNA are apparently located in some 40 different places in the *Drosophila* genome. The genes for the five major histones have been found, in several species, to be in tandemly repeated arrays. Moreover there are large non-coding regions between the different coding sequences, not all of which are read off the same chain. Clearly we shall learn much, in the next few years, about the distribution of the various kinds of sequences in eukaryotic genomes, especially in *Drosophila*, not only because of the concentration of work on this organism but also because its genetics will be needed to obtain decisive answers to our questions. The location of coding sequences and of single-copy non-coding sequences, the distribution of intermediate-repetitive sequences and in particular of the finely interspersed intermediate-repetitive sequences (strangely absent in *Drosophila*, though present in some of the larger diptera and in most higher organisms) should, hopefully, reveal some significant patterns.

It is one of our misfortunes that while we can with ease decipher coding sequences, we still have no reliable methods to spot promoters, terminators and operators, nor those sequences which may be needed for DNA domain formation and for RNA processing, let alone other instructions for functions as yet

This is an expanded version of the last part of the 8th Sir Hans Krebs lecture, entitled 'The structure of chromatin' delivered at the closing session of the 11th FEBS Meeting, Copenhagen, 19 August 1977.

Abbreviations. 3 D, 2 D and 1 D, three dimensional, two dimensional, one dimensional; mRNA, tRNA and hnRNA, messenger, transfer and heterologous nuclear ribonucleic acids; DNase, deoxyribonuclease.

unknown to us. We can hope for some progress in these areas from studies on the binding of the various non-histone proteins to DNA and on the mechanisms used for packaging and processing mRNA, a subject which at last seems to be starting to make some solid progress. Such experiments should be immensely helped by the high resolution 2D protein gels and the various affinity columns now coming into general use.

A major question is how much we can learn from small eukaryotic viruses, mainly oncogenic, such as SV40 (some of the latest work on this topic was described by Dr Paul Berg at the opening session). In small viruses a number of unexpected results have already been discovered. One of these is 'gene compaction' – the use of a single stretch of DNA to code for (part of) two quite different protein sequences, each read in a different phase. I suspect that this may occur more commonly in small viruses than in eukaryotic chromosomes. The fixed capsid size puts an upper limit on the amount of viral DNA that can be packaged, so it is not surprising that in such cases natural selection has had to make one DNA sequence do two jobs. I shall be surprised if this is often found in eukaryotic genomes themselves, at least in those of the higher eukaryotes, because there we have no obvious size limitations and, if anything, there appears to be an excess of DNA. If, in the course of evolution, one stretch of such DNA started to code for two distinct proteins I would expect gene duplication to occur, one copy then being evolved to code for only one of the proteins and the other copy to code for the other.

The fact that identical leader sequences are found in several of the late mRNAs for adenovirus, for example, although this leader is itself coded for elsewhere in the genome, may be due to the desirability of avoiding the repetition of DNA sequences in small viruses. The rate of recombination (per length of DNA) is so much higher for small viruses than for the host genome (in which repetitions are common) that repetitions in a viral genome may lead to an unacceptably high rate of deletion of the regions between them. The additional fact, that this leader sequence comes from not one but three distinct parts of the genome hints that, when put together, these parts produce an RNA sequence with a special tertiary structure. Such a compact structure may be needed to give extra stability to some of the late adeno mRNA molecules or may play some special role in the processing of the RNA. One would not be surprised if this tertiary structure (if it exists) turns out to be related to some known tertiary structure such as tRNA or its precursor.

These two novel features may be peculiar to small viruses. However, there is one aspect of these discoveries which may have a wider application. As has already been suggested by others, the fact that a single mRNA is coded by DNA in more than one place in the genome points to a novel and unsuspected mechanism for the processing of the nascent hnRNA. It had previously been assumed that the hnRNA was cut up into bits, some of which became mRNA, usually with the addition of a length of poly(A) at the 3′ end. The new alternative is that some of this processing is done by a looping-out mechanism, so that there is splicing as well as cutting of the relevant parts of the hnRNA. Such looping-out, cutting and splicing would allow the removal of unwanted sequences in the looped regions while bringing together those sequences which need to be made adjacent in the final mRNA. Different copies of the same type of hnRNA molecule may perhaps be looped out in different ways, thus producing different mRNAs, as required, from one type of hnRNA. The actual steps by which a composite messenger molecule is produced have not yet been established. Further work is needed to show whether these steps occur at the DNA level or the RNA level (or possibly both) and exactly how they are carried out.

These processes may well provide the missing clues needed to reveal the general structure of the eukaryotic genome. If the processes are at the RNA level, as seems probable, (except perhaps for special molecules like the immunoglobulins) then they raise the possibility that multiple promoters and operators may not be as common as has sometimes been supposed. More control may occur at the hnRNA level. Possibly DNA synthesis is needed to alter the packing of the chromosomal domains, so that certain changes in control at this level may only be possible during S phase, whereas control at the hnRNA level may occur at any time in interphase. Alternatively, changes to a domain may only be possible in prophase and may even require RNA synthesis. Clearly much more work is needed in this area.

If this looping-out mechanism for handling hnRNA proves true it would go a long way towards explaining the paradox of the high turn-over of that part of the hnRNA which never leaves the nucleus. The accumulation of extra DNA during evolution would than be seen as the consequence of mechanisms which multiply up existing stretches of DNA, distribute them rather randomly around the genome, where they can only be eliminated (should they not be needed) rather slowly, due to the low recombination rate. To work efficiently this process would appear to require an elimination mechanism at the hnRNA level somewhat related, in terms of the base sequences used as signals, to the postulated distribution mechanism at the DNA level. Organisms with long life-cycles (which tend to have large cells and large nuclei and which may therefore not be handicapped too much by an excess of DNA) might, by these methods, tend to acquire large amounts of DNA in their genomes.

Another suggestive line of work, favoured by some workers, is the so-called 'jumping-gene' phenomenon. My own view is that this probably does occur in higher organisms but that it will usually be rare, so that the process will be more important for evolution than for development but, here again, detailed evidence in eukaryotes, at the molecular level, is almost wholey lacking. One might suspect that such DNA shuffling is most likely to operate on the simple sequence DNA found in the various kinds of heterochromatin. It might well be at the bottom of position-effect-variegation.

If one stands back a little and tries to look at the picture as a whole, the most general unanswered question appears to be: how much does the 3D structure of the eukaryotic genome matter for expression, compared to the 1D structure? This is of great practical importance to the research worker, since 1D is so much easier to study than 3D. To find the 1D structure of any desired DNA sequence in, say, *Drosophila* is really only a matter of hard and careful work. This is mainly because we can produce fairly large amounts (by biochemical standards) of a "pure" gene as far as its DNA is concerned. This is not so easy at the 3D level. We have as yet, no method of obtaining, except in minute quantities, a pure undamaged 3D gene, protein and all. Unfortunately the prospect of reconstructing one accurately from its DNA and protein components does not, at this moment, look particularly rosy. In addition the methods of studying 3D structures with precision are far more difficult than the methods available for sequencing DNA.

Thus if it turns out that we can grasp the general nature of the eukaryotic genome purely from 1D studies we may hope for a relatively speedy answer. If, on the other hand, the 3D structure is not merely a packing device needed mainly for mitosis but is also of primary importance for gene expression, then the solution is likely to take longer and we will need a more devious and ingenious plan of attack.

Only time can show which alternative is preferred by nature and how difficult the problem will turn out to be. We certainly still have a long way to go but at least we can gain some comfort from the very large advances (at both the 3D and the 1D levels) which have taken place in the last three or four years.

The general nature of this talk makes the provision of detailed references difficult. The reader is referred to the papers in the chromatin and other sessions of this meeting and to the forthcoming volume based on the Cold Spring Harbor Symposium on 'Chromatin' held in June 1977.

F. H. C. Crick, Salk Institute for Biological Studies, P.O. Box 1809, San Siego, California, U.S.A. 92112